Synthese
https://doi.org/10.1007/s11229-019-02272-z

S.I.: DECTHEORY&FUTOFAI

A classification of Newcomb problems and decision theories

Kenny Easwaran[1]

Received: 24 March 2018 / Accepted: 28 May 2019

Abstract
Newcomb-like problems are classified by the payoff table of their act-state pairs, and the causal structure that gives rise to the act-state correlation. Decision theories are classified by the one or more points of intervention whose causal role is taken to be relevant to rationality in various problems. Some decision theories suggest an inherent conflict between different notions of rationality that are all relevant. Some issues with causal modeling raise problems for decision theories in the contexts where Newcomb problems arise.

Keywords Newcomb problem · Act-state dependence · Causal modeling · Decision theory · Rationality

1 Introduction

Consider the following two decision problems.

Example 1: The Newcomb Problem

You walk into a booth at a carnival and find a strange game being offered. There is a mad scientist, who offers you the option of taking just the contents of an opaque box, or the contents plus another box containing a thousand dollars. But as you walked in, her machines scanned your brain and body, and using the laws of physics and neurology, her computers predicted what you were going to choose. If they predicted that you would choose just the contents of the opaque box, she put a million dollars in it. But if they predicted that you would take both boxes, or if they couldn't predict what you would do, then she put nothing in. While you are deliberating, you learn that her predictions in the past have been moderately reliable, so that a significant majority (though not all) of her predictions of each type have been correct, conditioning either on her prediction

✉ Kenny Easwaran
easwaran@tamu.edu

[1] Texas A&M University, College Station, USA

or on the choice made by past participants. Should you just take the contents of the one opaque box, or take both boxes?

Example 2: The Smoking Lesion

Scientists discover that the statistician R.A. Fisher's work on behalf of the tobacco companies was actually correct. Although there is a strong correlation between smoking and lung cancer, it turns out that this is not because of any mechanism by means of which the nicotine, tar, or anything else affects human bodies. Rather, it is because of a previously unknown physiological condition that both has a tendency to cause cancerous growths in the lungs (and is the primary source of these growths), and seems to behaviorally lead to smoking. (This condition is caused by environmental factors that happen to have been cleaned up greatly in North America over the past few decades, but are still quite high in developing countries). You have had some cigarettes in the past, and have moderately enjoyed them. Should you take up smoking?

Classic "evidential decision theory" notes that in each case, the fact that an agent takes one of the options (taking both boxes, smoking) is good evidence that she will receive quite a bad outcome (failing to get her million dollars, cancer), and thus says that a rational agent who knows the setup will choose the other option. Classic "causal decision theory" notes that in each case, at the moment of decision, the features of the world that give rise to the bad outcome (prediction, physiological condition) are already fixed and outside of the agent's control, and instead focuses on the fact that the option that is correlated with this outcome will definitely cause a moderately good outcome (getting a bonus thousand, the enjoyment of a cigarette), and thus says that a rational agent who knows the setup will choose this option. Many people have expressed intuitions that cross-cut these cases, and thus some theorists have gone to great lengths to try to define versions of decision theory that recommend taking only one box in The Newcomb Problem, but smoking in The Smoking Lesion. (See, among others, the "ratifiability" of Jeffrey (1981), the "tickle defense" of Eells (1982), the attack on causal decision theory by Price (1986), the "cohesive decision theory" of Meacham (2010), and the "timeless decision theory" of Yudkowsky (2010)).

This paper does not seek to adjudicate between these theories, but instead attempts to give a structure for classifying problems of the relevant sort, and to use this structure to give some insight into the classification of possible decision theories for dealing with them. I classify the problems by means of a payoff table (which these two examples happen to share), together with a causal model explaining the correlation between the row and column of the payoff table (on which these two problems differ). I classify the decision theories by citing which particular variable in the causal model is the one that should be conditioned (evidential decision theory) or intervened on (the act itself for causal decision theory, and the decision procedure for timeless decision theory) in order to generate the probabilities used to evaluate the rows of the payoff table. (I consider decision theories as evaluative theories that classify certain acts as rational or irrational, rather than as procedures that agents themselves ought to use to generate their decisions).

These two classification systems then suggest some ways of generating new cases of the relevant sort (by combining some particular payoff table with a causal model), and new theories as well (by finding a different variable to condition or intervene on). I consider some motivations for thinking that one of these theories might be the correct account of decision-theoretic rationality, or whether all might be correct accounts of different parts of it. I close with discussion of some interactions of the concepts of free will, determinism, and rationality that are suggested by these classifications, though there are no definitive claims about these issues.

2 Classifying decision problems

One of the puzzling features of these cases is the fact that the "states of the world" are not probabilistically independent of the agent's choice. Thus, these decisions can't be properly analyzed by the methods proposed by Savage (1954) without having vastly more information than is available. (Savage requires a formulation of the problem where the states identified really are independent of the act chosen, which *might* involve "states" like "the predictor predicts correctly", or might involve some more subtle analysis. A proper Savage-style formulation of Newcomb-type problems is a major problem that is beyond the scope of this paper). Decision theorists dealing with Newcomb-type problems have more often tried to deal with them by giving some way to evaluate acts that are not always assumed to be independent of the "states".

For simplicity, I will focus on decision problems that are naturally described in terms of two possible acts and two possible states, which have some sort of correlation with each other. My proposed classification of these problems consists of two parts. First, I will characterize the value of the possible outcomes to the agent by giving a payoff table. Second, I will describe the structure of common cause relations giving rise to the correlation between the states and the acts. Although these parts leave out most of the real-world texture and background of the problems, I suspect that they are largely adequate for the purposes of decision theory. Importantly, although The Newcomb Problem and The Smoking Lesion have the same payoff table of possible outcomes, they will have different common cause structures, and thus will not be classified the same on my account, the way they are often treated in the literature. This will open space for decision theories that recommend one-boxing in The Newcomb Problem and smoking in The Smoking Lesion, even though other decision theories ignore the differences between the common cause structures.

2.1 Payoff tables

I represent the values of the possible outcomes for an agent in a 2×2 matrix like $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, which I call a "payoff table". Each row of the payoff table corresponds to one of the possible actions the agent can take and each column corresponds to one of the states. The contents of each cell then represents the value of the outcome that would be achieved by the corresponding combination of act and state. In most cases, the precise value of the outcome isn't relevant, so we will only be concerned with the ordinal structure of the value of these four outcomes. Thus, I will generally represent the values in the payoff table with the numerals 1–4, with 1 representing the best outcome and 4 the worst. When relevant, I will supplement this with some more detailed comparisons—for instance, $1 > 2 \gg 3 \sim 4$ will represent the fact that for the analysis of the relevant decision problem, it matters that the best option is better than the second best, which is much better than the third and fourth, which are approximately equal. When merely ordinal information is relevant, there are 24 possible payoff tables, though there are more if these magnitudes of comparison are considered.

Since switching the two rows just corresponds to relabeling the actions the agent can take, and switching the two columns just corresponds to relabeling the states, each decision problem has four possible payoff tables. I will cut this to two by following the convention that the act corresponding to the top row is positively probabilistically correlated with the state corresponding to the left column, and the act corresponding to the bottom row is positively correlated with the state corresponding to the right column. (When we look at the structure of the correlations in Sect. 2.2, it will be clear that there are further complications that I am ignoring right now). To remove this last degree of freedom, I will choose the representation where the upper left correlated outcome is strictly better than the lower right one. (If the upper left and lower right outcomes are equal, then I will ask that the upper right outcome is strictly better than the lower left one. If those are equal too, then both representations will in fact be identical). In addition to giving each decision problem a unique payoff table (at least, when the two correlated outcomes have distinct values), this eliminates half of the possible payoff tables, so that there are only 12 when merely ordinal information is represented.

The first thing to note is that both The Newcomb Problem and The Smoking Lesion are represented by the payoff table $\left(\begin{smallmatrix} 2 & 4 \\ 1 & 3 \end{smallmatrix}\right)$, with $1 > 2 \gg 3 > 4$. Each outcome of the lower action (taking two boxes, smoking) is slightly better than the outcome of the upper action (taking one box, not smoking), but the outcomes of the correlated left state (million dollars, no cancer) are much better than the outcomes of the right state (empty box, cancer). Several other puzzles that are prominent in the literature can be represented with the same payoff table.

Example 3: The Toxin Puzzle

(Kavka 1983, p. 33–34) You have just been approached by an eccentric billionaire who has offered you the following deal. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects. (Your spouse, a crack biochemist, confirms the properties of the toxin). The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you *intend* to drink the toxin tomorrow afternoon. He emphasizes that you need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives, if you succeed. (This is confirmed by your daughter, a lawyer, after she examines the legal and financial documents that the billionaire has signed). All you have to do is sign the agreement and then intend at midnight

> tonight to drink the stuff tomorrow afternoon. You are perfectly free to change your mind after receiving the money and not drink the toxin. (The presence or absence of the intention is to be determined by the latest 'mind-reading' brain scanner and computing device designed by the great Doctor X. As a cognitive scientist, materialist, and faithful former student of Doctor X, you have no doubt that the machine will correctly detect the presence or absence of the relevant intention).

Ignoring the central issue that Kavka uses this puzzle to discuss (namely, the question of whether it is even possible to form the relevant intention, knowing that your future self will have some incentive to go back on it), the value structure is quite similar. Regardless of whether or not you formed the intention tonight at midnight, not drinking the toxin tomorrow will be somewhat better than drinking the toxin. However, regardless of whether or not you drink the toxin, having formed the intention tonight results in a substantially better outcome than not having formed the intention. There is presumably some sort of correlation between having formed the intention and drinking the toxin, and it's better to have formed the intention and drink the toxin than to do neither.

Example 4: Parfit's Hitchhiker

> (Parfit 1984, p. 7) Suppose that I am driving at midnight through some desert. My car breaks down. You are a stranger and the only other driver near. I manage to stop you, and I offer you a great reward if you rescue me. I cannot reward you now, but I promise to do so when we reach my home. Suppose next that I am transparent, unable to deceive others. I cannot lie convincingly. Either a blush, or my tone of voice, always gives me away.

Here again we have an issue for planning. If I sincerely plan now to give you the reward when I get home, then regardless of whether or not I actually do give you the reward, I will get the major benefit of a ride out of the desert. However, actually giving you the reward is a cost to me, and it is correlated (in some sense to be discussed later) with the sincere plan. (In this case, if I don't plan to give you the reward, I may not even get the option to give it, because I'm stuck in the desert, so we have $3 \sim 4$ rather than $3 > 4$).

Example 5: Prisoner's Dilemma with a Twin

> You and your twin sibling have robbed a bank. However, you were caught and are now being held in separate cells. The police have enough evidence currently to convict each of you of conspiracy. However, they don't have enough evidence to convict either of you of actually robbing the bank. The police say that if you testify against your sibling (giving them enough evidence to convict), they'll drop your conspiracy charge. They're also making the same offer to your sibling in the other cell. Conspiracy carries a sentence of a year in jail, but bank robbery carries a sentence of five years.

In this case, regardless of whether or not your sibling testifies, if you testify, you'll spend one less year in jail. However, because your twin is very much like you, your testifying is strongly correlated with your twin testifying, which would add five years to your jail sentence whether you testify or not.

Further analysis of the similarities and differences among these decision puzzles will wait until Sect. 2.2.

The distinct, but related, payoff table $\left(\begin{smallmatrix} 3 & 2 \\ 1 & 4 \end{smallmatrix}\right)$ also corresponds to a few prominent puzzles in the literature.

Example 6: Death in Damascus

> (Gibbard and Harper 1978, pp. 185–186) Death works from an appointment book which states time and place; a person dies if and only if the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of highly reliable predictions. An appointment on the next day has been inscribed for him. Suppose, on this basis, the man would take his being in Damascus the next day as strong evidence that his appointment with death is in Damascus, and would take his being in Aleppo the next day as strong evidence that his appointment is in Aleppo. Two acts are open to him: A, go to Aleppo, and D, stay in Damascus. There are two possibilities: SA, death will seek him in Aleppo, and SD, death will seek him in Damascus. He knows that death will find him if and only if death looks for him in the right city.

In this case, the rows correspond to the actions of going to Aleppo and staying in Damascus. Two states of the world are possible, death seeking him in Aleppo and death seeking him in Damascus. The best outcomes (1–2) occur if the man goes where death isn't seeking him, while the worst (3–4) occur if he goes where death is seeking him.

Example 7: The Psychopath Button

> (Egan 2007, p. 97) Paul is debating whether to press the "kill all psychopaths" button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world *with* psychopaths to dying.

In this case, the upper row is the act of not pressing the button while the lower row is pressing, and the left column is the state of not being a psychopath while the right column is being one. The best outcome is pressing the button while not being a psychopath, and the worst is pressing while being a psychopath. Not pressing is in between, so we have $1 > 2 \sim 3 \gg 4$.

Both of these puzzles have been proposed to cause various problems for causal decision theory. Causal decision theory leads to a kind of instability in Death in Damascus (whatever decision one makes, one will then think that the other would have been better). But it actually seems to lead to the intuitively wrong recommendation in

The Psychopath Button (if Paul's prior probability that he is a psychopath is relatively low, then causal decision theory tells him to press the button, which seems like a really bad idea). The important difference here is primarily due to the fact that for Death in Damascus the ordinal structure is really $1 \sim 2 \gg 3 \sim 4$, while for The Psychopath Button the ordinal structure is really $1 > 2 \sim 3 \gg 4$. (One might prefer to represent Death in Damascus with $\binom{2\,1}{1\,2}$ and The Psychopath Button with $\binom{2\,2}{1\,3}$, but I will continue to use four distinct numbers rather than expanding my search to payoff tables with two or three distinct entries as well).

Some of the other possible payoff tables have been investigated in the context of game theory, though not so much in the context of evidential or causal decision theories. In the game theory literature, $\binom{1\,4}{2\,3}$ corresponds to the "Stag Hunt", while $\binom{1\,3}{4\,2}$ corresponds to a "coordination game".

A few other payoff tables are unlikely to yield any interesting problems. With $\binom{1\,2}{3\,4}$ and $\binom{1\,3}{2\,4}$, the top action is preferred to the bottom one regardless of state, and the left state is preferred to the right one regardless of act. In the payoff table for The Newcomb Problem, it is also the case that one act is preferred to the other regardless of state, and one state is preferred to the other regardless of act, but the interest arose because the *preferred* act was correlated with the *dispreferred* state. These two are uninteresting because the preferred act and preferred state are correlated, so it is hard to imagine a plausible decision theory recommending anything other than this preferred act.

However, cases with the table $\binom{2\,1}{4\,3}$ may yield some interesting considerations, if $1 \gg 2 > 3 \gg 4$. Even though there is a sort of "superdominance" of the top action over the bottom action (in that any outcome of the top action is better than any outcome of the bottom action), the strong interest in getting the much better outcomes of the right rather than left column may motivate some attempt to take advantage of the correlation of the right column with the bottom row, even at the cost of running the risk of actually choosing the bottom action. Describing a thought experiment that motivates this reasoning may be an interesting project for future work.

2.2 Causal models

Having considered the payoff tables, it is useful to investigate the nature of the correlation between act and state. My overarching methodological assumption is a version of Reichenbach's Common Cause Principle—if two things are correlated, then either one must be a cause of the other, or there must be some common cause. (See Arntzenius (2010) for careful formulations of the principle). In The Newcomb Problem, the correlation between how many boxes the agent chooses and how much money is in the boxes is because the brain state of the agent when she walks into the carnival tent both causes the scientist to distribute money one way and causes the agent to decide one way. In The Smoking Lesion, the correlation between cancer and smoking is because both tend to be caused by the presence of a physiological condition and prevented by the absence of the physiological condition. I mean to rule out any proposed Newcomb-type problem on which there is no such explanation of the correlation.

My approach is to analyze the causal structure of the situation using one of the "causal modeling" frameworks of Pearl (2000) or Spirtes et al. (2000). The differences between these frameworks might be relevant for detailed analysis of some cases, but for present purposes, an outline of the shared idea may be sufficient.

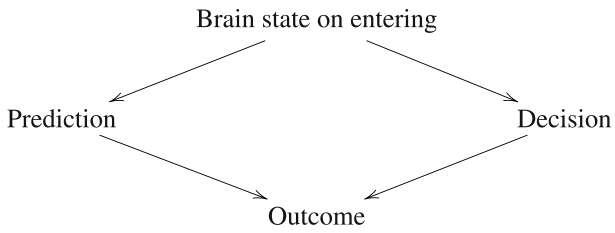
A causal model represents the causal relations among various variables of interest. These variables might be binary (the money is in the box or not; the agent takes one box or two) or might have multiple possible values (the amount of fertilizer added to a field; the height of the crops in the field). Each variable is represented by one node in a network, with an arrow going from one variable to another if the values of the first play a direct role in determining the values of the other. (We label the first as the "parent" and the second as the "child"). A central assumption is that causation is non-circular—there is no path from a variable to itself that follows arrows only in the direction they point. Thus, for any two variables, there is a unique fact of whether the first is a causal ancestor of the second, the second is a causal ancestor of the first, or neither.

With a given network, the model can be specified by providing, for each variable, a probability distribution over all its values conditional on each particular setting of values for its parents. (For variables with no parents, this is just one unconditional probability distribution). These conditional probability distributions for each variable define a joint probability distribution for all the values of all the variables collectively—the probability of any particular setting of all the variables in the network is equal to the product of the probability of the specified value of each variable conditional on the specified values of its parents. This joint probability distribution then satisfies Reichenbach's Common Cause Principle with the causal relations specified by the network—two variables will have a probabilistic correlation only if one is a causal ancestor of the other, or they share a causal ancestor in common. At the level of analysis I'll be using, it won't be essential to give the detailed probability distributions—I'll just provide the network diagrams. But to give a properly detailed decision-theoretic analysis, the specific probabilities will be needed (just as the specific utility values for outcomes will too).

Importantly, in addition to defining a joint probability distribution over the full set of variables, and probability distributions for each variable *conditional* on others, these models define probabilities given an *intervention* on one of the variables to set it to a particular value. To calculate these probabilities, we consider a related causal model in which the probability distribution for that variable is set to probability 1 of having that particular value, regardless of the values of its parents, and then calculate probabilities in the new model. Although the probabilities of a variable *conditional* on values of its children are generally different from their unconditional probabilities, the probabilities of a variable when *intervening* on one of its children are equal to the unconditional probabilities.

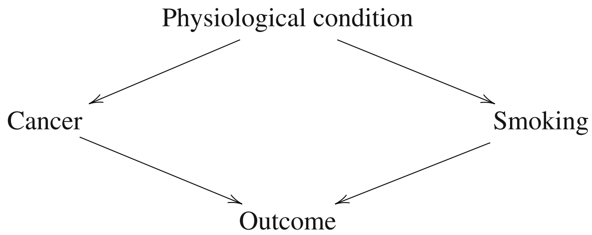
My first attempts at causal models for these problems will represent the causal structure of the world only partially. Only variables that are explicitly mentioned in the statement of the problem will be mentioned, and single arrows may hide a whole chain of intermediate causal variables that weren't explicitly mentioned. As a result, many of these cases will at first have structures that look the same. But for my classification purposes, it will be important later to expand some of these arrows, to make sure that certain important variables are explicitly mentioned, rather than being implicit within the chain of actual events that constitute the causal connection represented by a single arrow. Some decision theories will say that it matters where exactly in this chain the common cause diverges.

For The Newcomb Problem, a simple causal network diagram might look something like this:



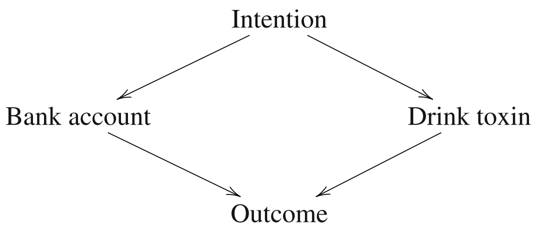
The outcome of the problem is most directly determined by whether or not the predictor put the money in, as well as the agent's decision whether to take one or two boxes. The prediction and the decision are correlated, because both are in some sense determined by the agent's state on entering the booth. This determination may be probabilistic—perhaps the brain states that the predictor is able to read merely make certain decisions on the part of the agent very likely, or perhaps the predictor's scanner is not 100% accurate.

For The Smoking Lesion, the initial diagram is similar:

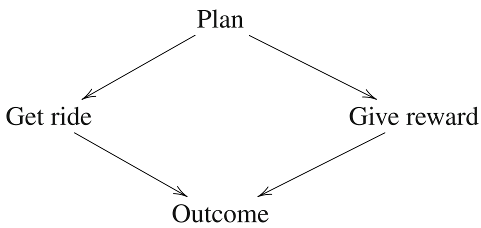


In this case, "Outcome" may not be a distinct event that is *caused* by the fact of whether the agent has cancer or not and smokes or not, but may rather be *constituted* by it. But in order to see how the overall satisfaction of values for the agent is brought about by the various factors involved, we need a single node to represent it.

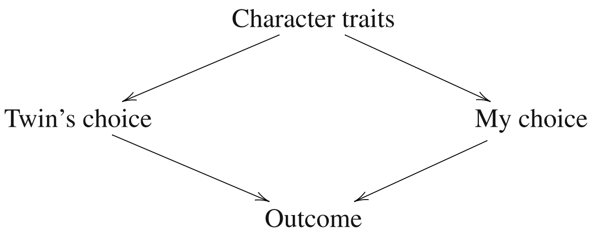
The diagrams are also very similar for The Toxin Puzzle:



and Parfit's Hitchhiker:

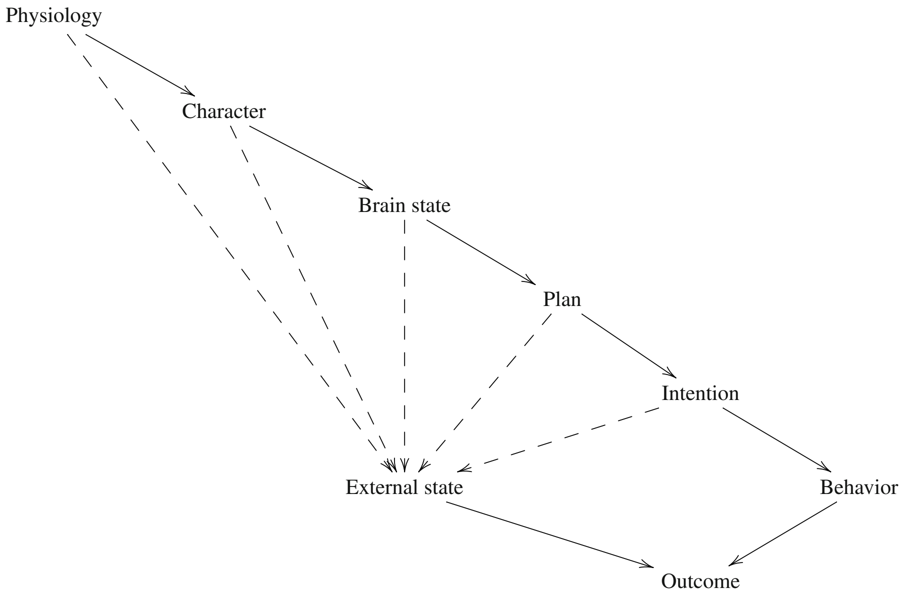


and Prisoner's Dilemma with a Twin:

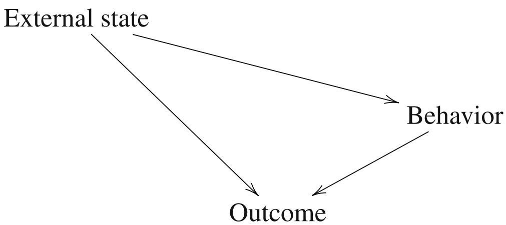


All of these puzzles have been considered Newcomb-like, because all involve some common cause of the external state and the agent's choice, inducing the correlation that underlies the problem. When we just include the one common cause, the diagrams look structurally identical.

However, some decision theorists have noted that these similar structures hide important differences between the problems. Although each of the common cause variables is only labeled in one of the problems, character traits, brain states, intentions, and plans are part of the causal chain in every one of the problems, with the relevant difference being at what point the cause of the "external state" variable branches off from this chain. All of the decision problems have the solid arrows in the following diagram (as well as additional incoming arrows), but each has a different one of the dashed arrows:



Once we consider this causal model, we can develop other cases where the correlation takes a different form. For instance, one might take the external state to be correlated with the behavior not because of a common cause, but because the external state is itself a cause of the behavior.



This is a plausible first pass at the structure in The Psychopath Button (the more detailed account might identify the "external state" with "physiology" or "character" or "brain state", or might consider it as a partial cause of one of those), and also in the following case from Christopher Hitchcock, whose payoff table is more like the original Newcomb case:

Example 8: Banana

(Hitchcock 2016) Suppose that on some mornings, you eat a banana. You don't put any thought into this, nor do you experience any phenomenologically accessible banana-cravings. Your mornings run on auto-pilot, and sometimes you just find yourself eating a banana.

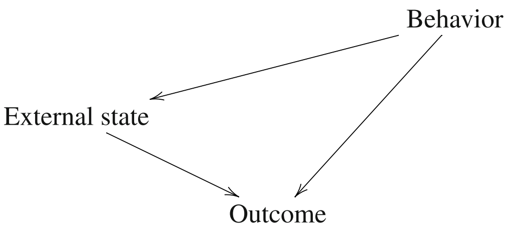
> You notice that on days when you eat a banana, you often suffer from a painful, migraine-like headache. You occasionally suffer from these headaches when you don't eat a banana, and on those days, the headache is even worse.
>
> You believe the following about the causal structure of this phenomenon: 10% of the time, you wake up suffering from a potassium deficiency. …You are much more likely to eat a banana on days when you suffer from a potassium deficiency. …If you suffer from a potassium deficiency, you will suffer from a headache of intensity 10. Eating a banana reduces the intensity by one, and no headache occurs if there is no potassium deficiency.

For that matter, there can also be cases where the correlation arises because the behavior is itself directly a cause of the "external state".

Example 9: Study or Party

> There will be a big test tomorrow worth a large fraction of the grade in a class that you are taking. Some of your friends are also throwing a party tonight. Regardless of whether you pass or fail the test, you would like to go to the party. But regardless of whether or not you go to the party, you would *really* like to pass the test. You suspect that going to the party tonight, rather than studying will cause you to have a substantially lower probability of passing the test.

In this case, the causal network is something like the following, where "behavior" is either studying or partying, and "external state" is either passing or failing the test.



More complex cases might involve multiple causal arrows that all partly contribute to the correlation. We could imagine a version of The Toxin Puzzle in which brain scans are conducted at several points, and for each one that points towards drinking the toxin, a coin is flipped, and the money is placed in the account if *any* flip comes up heads. There could also be versions where some branches of the causal diagram correlate the state and act variables while other branches tend to anti-correlate them. (This is perhaps a way to explain what is going on in the "Meta-Newcomb" puzzle of Bostrom (2001)). As noted by Cartwright (1979), some of these cases might even involve no apparent correlation between act and state, because multiple causal pathways could cancel out (for instance, the extent to which bananas prevent headaches might exactly counteract the correlation with headaches produced by the potential common cause of potassium deficiency). There is an interesting distinction between these cases (where act and state have multiple causal pathways connecting them that cancel out) and cases in which there are no causal pathways between the acts and states. The former are violations of the "faithfulness" assumption of Spirtes et al. (2000), and thus are expected to be quite rare (because usually the multiple causal pathways will fail to exactly cancel out in their influence). (I thank Reuben Stern for this point).

Thus, I classify a Newcomb-type problem by providing a payoff table and a causal model. The payoff table should specify which variables in the causal model correspond to the rows and columns of the table. The causal model should explicitly include not just the row and column variables, but also any common causes that give rise to the correlation between them. Furthermore, in order for various decision theories described in the next section to be able to operate with these classifications, the causal model should also make clear the position of each psychological variable on which one might conceivably intervene.

Combining a payoff table and a causal model, we can generate further cases with aspects of the Newcomb structure. Some of these cases may generate different intuitions than the ones considered so far, and some may be useful for testing various decision theories. So at this point it will be useful to turn to the theories.

3 Classifying decision theories

So far as I know, every proposed decision theory for problems of this sort evaluates the rows of a payoff table by means of a calculation of a kind of "expected value". Each row is evaluated by summing the values that occur in each cell of that row, multiplied by some probability of the column that the payoff occurs in. (In this section I will refer to these as "column" and "row" rather than "state" and "act", because in some cases there are multiple variables that might all be said to represent "states" and perhaps multiple variables that represent "acts"). But the decision theories differ based on *which* probabilities are considered relevant. When we have classified the decision problem by means of the payoff table and the causal model, it is straightforward to see how the payoff table must figure into this calculation. But the question is how to use the causal model to determine the relevant probabilities. We can classify a decision theory by means of the way it extracts the probabilities relevant to evaluating rows from the causal model that represents the correlation between rows and columns.

As mentioned above, a causal model determines unconditional probabilities for every variable in the structure, and it also determines probabilities for each variable *conditional* on any particular value of another variable, as well as probabilities for each variable for an *intervention* setting any other variable to a particular value. A very simple decision theory is one that calculates expected value of a row by using the unconditional probabilities of the columns determined by the causal model. This sort of view is sometimes attributed to Savage (1954), though he specifically didn't consider problems where the rows and columns of the decision table are correlated. As far as I know, no one actually defends this decision theory as a general guide for problems of this type, because it says to choose the row with higher payoffs in every problem with the same payoff structure as The Newcomb Problem—but in Study or Party this means partying, which seems blatantly wrong. (Because the unconditional probabilities are the same in calculating the expected value of each row, the row with higher payoffs ends up having a higher expected value than the row with lower payoffs, even if intuitively the columns should be weighted differently for the two rows).

To fix this, evidential decision theory, as advocated by Jeffrey (1965), evaluates each row using the probabilities of the columns after *conditionalizing* on that value of the row variable. Enough of the time, when you study you pass the test, and when you party you fail the test, so studying is better than partying. But some theorists have felt that this gives unintuitive responses to Banana or The Newcomb Problem or The Smoking Lesion, where it recommends not eating the banana, one-boxing, and not smoking, since being in these rows of the decision table is strongly enough correlated with being in the good column.

Causal decision theory, as advocated by Joyce (1999), advocates considering the probabilities given by the structure when one *intervenes* on the row variable to set it to that value. When the only parent variables of the column variable are the row variable or its causal descendants, as in Study or Party, this agrees with evidential decision theory. But when the variables bear some other relationship, the decision theories can disagree. Causal decision theory recommends eating the banana, two-boxing, and smoking, but without recommending partying.

At the beginning, I mentioned that many theorists are interested in a theory that would recommend one-boxing in The Newcomb Problem and smoking in The Smoking Lesion. Such a theory can't operate using probabilities derived from the unconditional causal model, or from the causal model with conditionalization or intervention on the variable corresponding to the row. However, if the relevant probabilities are ones derived from conditioning or intervening on some *other* variable in the structure, this is possible.

The motivation for considering probabilities given by an intervention on the causal model, rather than conditionalization, is, as suggested by Cartwright (1979) and Hitchcock (2016), because intervening is associated with the *effectiveness* of bringing about an end. However, Hitchcock notes that there is a controversy over whether interventions really are *possible* at the level of behavior, particularly in Newcomb-type cases where the causal model already specifies the way psychological states like intention and decision factor into the decision. As Yudkowsky and Soares (2018) put it, "[causal decision theory] agents fail because they imagine breaking too many correlations between action and environment." It could be that it is best to think of intervention at some other variable in the model, or in terms of conditioning on one of the relevant variables to generate probabilities, instead of intervening.

Once we have looked at the structures underlying the various cases described above, we see a range of variables that could all be plausible sites of conditioning or intervention in the larger causal model. A person's character, brain state, plans, intentions, and behaviors are all to one degree or another at least partially under her control. However, in many cases, the control an agent has of one of these features goes by means of her control of the others.

Furthermore, the control an agent has of any of these features is often quite far from perfect. Much literature in epistemology questions the extent to which decision-theoretic models could be appropriate for studying belief, because belief is in some sense not under our voluntary control. Consideration of these Newcomb-type cases suggests that perhaps *action* in some ordinary sense isn't really under voluntary control either, at least not in a situation where the environment already encodes determined features of our behavior. We might ask what *sort* of voluntary control is needed for our actions to be subject to normative theories, and do we have this sort of control over actions, intentions, plans, mental states, or any of the other possibly relevant features? Hieronymi (2006) argues that in fact beliefs and intentions are both out of our control in the same sort of way (distinct from the control that we have over our actions), but she also argues (2008) that this sort of lack of voluntariness is a hallmark of things for which we bear a certain sort of normative responsibility. We control our beliefs and our intentions only *by* directly reasoning about the facts or actions that they are supposed to correspond to in the world. I don't know if this is the right way to think about the sort of control we have over our moral or practical character, or virtues, or other features that are relevant for some of the problem cases. But there is some notion of control here and this might suggest multiple points of intervention in these problems.

Once we consider personality, character, mental state, intention, plan, and action all to be sites of potential intervention, there's a question about which, if any, of these is the proper locus of rationality. Causal decision theory says it is the furthest downstream site of possible intervention—at the row variable, corresponding to the behavior itself. But a decision theory could calculate probabilities on interventions farther upstream. Perhaps the right level to analyze is interventions on plans, or intentions, rather than actions. Perhaps the right level is interventions on character. As long as the intervention is somewhere downstream of one's physiology and upstream of one's actions, the theory will recommend smoking in The Smoking Lesion, and studying in Study or Party. But for the other decision problems, what rationality recommends depends on whether the intervention that is relevant is upstream or downstream from the point of common cause that establishes the correlation.

My classification of decision theories is given by the way the causal models determine the probabilities they use—is it by conditioning or intervening, and which particular variable in the causal model is the target variable? No particular decision theory cares about the whole causal model—they just care about whether the correlation between act and state is determined upstream or downstream from the relevant point at which probabilities are calculated. But I need the causal model involved in classifying the decision problem to be quite detailed since each particular decision theory will calculate probabilities using a different part of the model.

3.1 Unitary decision theories

Rationality is a concept that appears to apply to many different sorts of bearers—actions, plans, character trains, etc. A very plausible-seeming "enkrasia" principle is the following. It is rational to *plan* to do X iff it is rational to *do* X. Similarly, one might say that the rational *character traits* to have are the traits that lead one to do rational *acts*. The rational kind of *mental state* to have is the kind that leads one to make rational *plans*. This sort of principle speaks to a unity of rationality. In order to figure out what rationality means in general, we need to figure out which site of intervention (or conditionalization) is fundamental, and rationality for the others follows from that. Traditional classroom presentations of normative ethical theories often take this kind of form as well—either outcomes (consequentialism), acts (deontology), or personal character traits (virtue ethics) are said to be the fundamental site of moral evaluation, with the evaluation of the others determined by seeing which traits, acts, and consequences are linked to each other.

Traditional causal decision theory is often presented in a way that presupposes that the moment of action is the primary site for rational evaluation, and that this is where intervention is applied to the model. Plans and character traits and so on are then evaluated by seeing what momentary actions they lead to. Two-boxing is the rational act in The Newcomb Problem, because an intervention at the last moment to two-box would result in the best outcome. Since two-boxing is rational, this means that the rational plan to make is to two-box, and a rational person is one who two-boxes, and so on. The primary bearer of rationality is the act. The intention, the psychology, the character, etc., are only rational in a sense that is parasitic on the acts that they tend to produce.

There is a kind of tragedy of rationality for this view. Rational people end up with only $1,000 in The Newcomb Problem, and they end up in a situation of mutual defection in Prisoner's Dilemma with a Twin. They can't get the prize in The Toxin Puzzle and they get abandoned in the desert in Parfit's Hitchhiker. Rationality is still right, even if irrational people (according to the traditional causal decision theorist) often end up much better off than the rational ones do. Lewis (1981) argues that this tragedy is just part of life in an imperfect world. Sometimes the world rewards the irrational, but that doesn't mean that they're rational. However, some theorists, like Greene (2019), argue that this tragedy should lead us to reconsider our decision theory.

If the moment of action were the only conceivable point of intervention, this traditional causal decision theory would seem very plausible. But other points of intervention are possible. Many of our individual behaviors seem to be best controlled by means of developing habits, or practicing, or shaping our character, rather than by trying really hard to be rational at the moment of action. Thus, some theorists have taken this tragedy to be a sign that traditional causal decision theory has things backwards. Rather than evaluating character traits, mental states, intentions, plans, acts, and other points of intervention by means of probabilities derived from intervening on the one of those variables that is farthest *downstream* in the causal chain, we might evaluate them by means of probabilities derived from intervening on the one that is farthest *upstream* in the causal chain. Decision theory, on this picture, is about what kind of person to be, rather than what to do in the moment.

This is often cashed out through a metaphor of taking the goal of rationality to be the design of the best overall algorithm for reacting to decision problems, treating the self almost as a robot to be pre-programmed. (Yudkowsky 2010; Yudkowsky and Soares 2018) Where the traditional causal decision theorist says that rationality is about being the kind of person that does in the moment the most successful thing to do in that moment, this sort of view says that rationality is about doing in the moment what the most successful kind of person does. One should be the kind of person that gets the prize if one can, even if that means leaving money on the table (literally, in the case of The Newcomb Problem). If one rationally formed plans or intentions or character traits in the past, one should go along with them in the present. Even if one *hadn't* formed these plans or intentions (perhaps one finds oneself in a situation analogous to one of the ones above, without any advance warning), one should act *as if* one had. Where the causal decision theorist uses probabilities given by the farthest downstream possible intervention, Yudkowsky uses probabilities given by the farthest *upstream* possible intervention.[1]

This sort of impulse exists in many related philosophical domains. Process reliabilism (Goldman 1979), virtue epistemology (Turri et al. 2017), Kantian ethics, and rule utilitarianism (Hooker 2015) all take this kind of form. A belief is evaluated for rationality not by evaluating the belief itself for truth, but instead by evaluating the process or epistemic virtues that gave rise to it for their truth-aptness. An act is evaluated for morality not by evaluating its own consequences, but by evaluating the consequences of the maxim or rule that gave rise to it when considered as a guide in general. Although it is still the consequences that guide the evaluation, an act or belief is evaluated not by its own consequences, but by the consequences of some causally upstream feature that gave rise to it.

Causal Decision Theory and Yudkowsky's Functional Decision Theory thus represent two natural endpoints for enkratic decision theories that focus on one primary point of evaluation. They choose either the farthest downstream or the farthest upstream point of evaluation and compare options by expected value in the probability distribution determined by intervening on that point. All other points of evaluation are then evaluated based on which one of those options they connect up with. Other theories could choose points in between as the most natural point of intervention. And of course, like Evidential Decision Theory, one could choose to evaluate by conditionalizing instead of intervening on the relevant variable.

3.2 Pluralist decision theory

However, it is not obvious that this sort of enkratic principle is required. It seems to ignore the fact that there are *many* possible points of intervention, and that even the person who has made the right intervention at one point (to develop a habit, to consider a maxim, to form an intention, to devise a rule) could still do better by *also* intervening at another (to act in a way other than the one normally caused by that habit, maxim, intention, or rule). The kind of "global consequentialism" discussed by Pettit and Smith (2000) and Kagan (2000) says that we should evaluate each of these "evaluative focal points" on the basis of the consequences of intervening on *it* rather than others. Streumer (2003) and Lang (2004) object to this global consequentialism by noting the tension between the recommendations—a general inclination to intervene at the further downstream points would itself be a character trait that contradicts the upstream trait. But Driver (2014) suggests that "judging a normative ambivalence to

[1] His version of the theory has some additional complexities because he characterizes agents as deterministic algorithms, so that the intervention must be in some sense logically impossible. In a case like Prisoner's Dilemma with a Twin, the evaluation works by assuming that I and the twin both implement some particular deterministic algorithm X. Then we note, if algorithm X were to output cooperation we would both do pretty well, and if algorithm X were to output defection we would both do pretty badly, so it ought to output cooperation. Strictly speaking, at least one of those counterfactuals will be counter-*logical*, if an algorithm has its outputs as a matter of logical certainty.

be *appropriate* rather than problematic goes a long way toward mitigating the air of paradox." A decision theorist might say there is just a pluralism of rationalities. There is one notion of rational act, and another notion of rational plan, and another notion of rational character.

This pluralist way sees a kind of dilemma, or even an essential tragedy, for rationality. The rational act may not be the one that would be performed by an agent with the rational character. Character traits that lead one to pick one box in The Newcomb Problem, pay up in Parfit's Hitchhiker, and drink the toxin in The Toxin Puzzle are rational. But rational acts are the ones that go against these character traits. There is one question of what kind of person to be, and another question of what to do, and the answers to these questions just unfortunately don't cohere in Newcomb-like cases. If we had perfect control over momentary actions, then maybe we should agree with the traditional causal decision theorist about rationality—but then there would be no such thing as general character traits that determine our actions, and that we could shape. If we had perfect control over the kind of person we are, then maybe we should agree with Yudkowsky—but these traits would take away the possibility of direct control of action. These problems arise only when there are multiple points that all have some claim of being a point at which we can intervene.

4 Bigger questions

The viewpoint I have described on Newcomb-type problems categorizes them by the payoff table of potential outcomes, and the causal model giving the correlation between act and state. It categorizes the decision theories based on which theoretically possible point(s) of intervention or conditionalization generate the probabilities used to rationally evaluate rows of the decision table.

I have not addressed the question of how to deal with uncertainty about the causal model. (Most traditional problems stipulate that the relevant agent knows for sure what is going on). And I have not addressed the question of how to figure out what the causal model is. But this is a major part of the general problem of induction, and it is a lot to ask of a decision theory to answer that question!

The various understandings of "intervention" in these causal models raises some further questions for the notion of rationality in these cases. On the picture described above (often associated with Pearl (2000)), intervention involves a change in the causal model, and the act of intervention must come from outside the model and break the structure that exists. This is a natural point of view for models that are inherently incomplete, representing the causal structure of only part of the world. This is often appropriate for an understanding of experimental design, where the experimenter can be treated as an aspect of the universe that is outside the scope of the model, and can break its relations in some cases. But it may not be appropriate for a theory that is supposed to tell us something about the normative character of rationality, particularly if the cases (like The Newcomb Problem, The Toxin Puzzle, Prisoner's Dilemma with a Twin, and even Banana) are ones where it is assumed that *all* causal variables are relevant, *including* ones that affect the outcome at least in part by determining the mental states of the agent.

On another picture (often associated with Spirtes et al. (2000)), an intervention is just itself another variable within the causal model, and it counts as an intervention only to the extent that one setting of this variable approximately screens off its child from its other parents. (*Conditioning* on this setting of the intervention variable approximates the probability distribution given by *intervening* on the child variable on Pearl's picture). However, if we get a complete causal model of the world, it seems likely that no variable will *perfectly* screen off its children, and conditioning on such a node will only at best *approximately* represent what we think of as an intervention. When we combine this with the fact that none of the sites of intervention are totally effective (particularly in Newcomb-like problems), it looks like some of the foundations of a traditional conception of rationality start to crumble.

On either picture, the apparent adequacy of a conception of intervention requires the model to be incomplete. But a truly normative picture of rationality that can deal with cases stipulated as precisely as some philosophers like to discuss, seems like it should work best when we have a complete model of the world. If this causal modeling structure is the right way to proceed, then the Newcomb-type problems may point to an essential incompleteness or approximation in our concept of rationality.[2]

References

Arntzenius, F. (2010). Reichenbach's common cause principle. *Stanford Encyclopedia of Philosophy*.

Bostrom, N. (2001). The meta-Newcomb problem. *Analysis*, *61*(4), 309–310.

Cartwright, N. (1979). Causal laws and effective strategies. *Noûs*, *13*(4), 419–437.

Driver, J. (2014). Global utilitarianism. In B. Eggleston & D. Miller (Eds.), *The Cambridge companion to utilitarianism*. Cambridge: Cambridge University Press.

Eells, E. (1982). *Rational decision and causality*. Cambridge: Cambridge University Press.

Egan, A. (2007). Some counterexamples to causal decision theory. *Philosophical Review*, *116*(1), 93–114.

Gibbard, A., & Harper, W. (1978). Counterfactuals and two kinds of expected utility. In W. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 153–190). Dordrecht: Reidel.

Goldman, A. (1979). What is justified belief? In G. S. Pappas (Ed.), *Justification and knowledge* (pp. 1–25). Dordrecht: Riedel.

Greene, P. (2019). Success-first decision theories. In A. Ahmed (Ed.), *Newcomb's problem* (pp. 115–137). Cambridge: Cambridge University Press.

Hieronymi, P. (2006). Controlling attitudes. *Pacific Philosophical Quarterly*, *87*(1), 45–74.

Hieronymi, P. (2008). Responsibility for believing. *Synthese*, *161*(3), 357–373.

Hitchcock, C. (2016). Conditioning, intervening, and decision. *Synthese*, *193*(4), 1157–1176.

Hooker, B. (2015). Rule consequentialism. *Stanford Encyclopedia of Philosophy*.

Jeffrey, R. (1965). *The logic of decision*. New York: McGraw-Hill.

Jeffrey, R. (1981). The logic of decision defended. *Synthese*, *48*(3), 473–492.

Joyce, J. (1999). *The foundations of causal decision theory*. Cambridge: Cambridge University Press.

Kagan, S. (2000). Evaluative focal points. In B. Hooker, E. Mason, & D. Miller (Eds.), *Morality, rules, and consequences*. Edinburgh: Edinburgh University Press.

Kavka, G. (1983). The toxin puzzle. *Analysis*, *43*, 33–36.

Lang, G. (2004). A dilemma for objective act-utilitarianism. *Philosophy, Politics, and Economics*, *3*(2), 221–239.

[2] The idea for this paper was originally developed at a workshop at the Machine Intelligence Research Institute in Berkeley in September, 2013. The paper was presented at the workshop, "Self Prediction in Decision Theory and AI" at Cambridge University in May, 2015, as well as at the University of Houston philosophy department colloquium in October, 2015, and at the Texas A&M University Workshop in Decision Theory and Epistemology in December, 2015. I also discussed the idea of pluralist decision theory on Julia Galef's "Rationally Speaking" podcast on August 9, 2015.

Lewis, D. (1981). Why ain'cha rich? *Noûs*, *15*(3), 377–380.
Meacham, C. (2010). Binding and its consequences. *Philosophical Studies*, *149*(1), 49–71.
Parfit, D. (1984). *Reasons and persons*. Oxford: Oxford University Press.
Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
Pettit, P., & Smith, M. (2000). Global consequentialism. In B. Hooker, E. Mason, & D. Miller (Eds.), *Morality, rules, and consequences*. Edinburgh: Edinburgh University Press.
Price, H. (1986). Against causal decision theory. *Synthese*, *67*, 195–212.
Savage, L. J. (1954). *The foundations of statistics*. New York: Dover.
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction and search*. Cambridge: MIT Press.
Streumer, B. (2003). Can consequentialism cover everything? *Utilitas*, *15*, 237–247.
Turri, J., Alfano, M., & Greco, J. (2017). Virtue epistemology. *Stanford Encyclopedia of Philosophy*.
Yudkowsky, E. (2010). *Timeless decision theory*. Technical report, The Singularity Institute.
Yudkowsky, E. & Soares, N. (2018). Functional decision theory: a new theory of instrumental rationality. Technical report, Machine Intelligence Research Institute.